Journal of the Korea Academia-Industrial cooperation Society
Vol. 25, No. 3 pp. 10-15, 2024

https://doi.org/10.5762/KAIS.2024.25.3.10
ISSN 1975-4701 / eISSN 2288-4688

GPT 4.0의 우회 공격과 LLM 침입 차단 모델에 대한 연구

이용준[1], 강장묵[1,2*], 김지효[1]
[1]극동대학교 해킹보안학과, [2]동국대학교 정보보호학과

Research on evasion attacks and LLM intrusion prevention models in GPT 4.0

Yong-Jun Lee[1], Jang-Mook Kang[1,2*], Ji-Hyo Kim[1]
[1]Division of Hacking Security, Far East University
[2]Department of Information Protection, Dongguk University

요 약 2023년의 1년은 인공지능 기술 발전 역사 50년보다 더 많은 변화가 있었다. 본 연구는 GPT-4.0의 주요 특징을 멀티 모달 기능, 언어 이해 및 처리, 성능, 정확성으로 분류하였다. GPT-4.0의 5가지 요인을 중심으로 거대언어모델의 취약점을 예방하고 해커 등의 침투 공격을 차단하는 기술/사회공학적 거버넌스 모델을 제시하였다. 본 연구는 거대언어모델의 탈옥 사례를 대상으로 하였다. Bard, open-AI, MS-bing 등에서 개별 침투 공격과 탈옥 사이트를 분석하였다. 해당 연구 대상 소스를 얻은 사이트는 제일브레이크 챗(https://www.jailbreakchat.com/)이다. 연구결과, 인공지능은 목적함수(Object Function)를 사용하여 최적화하기 때문에 목표(goal)을 하이제킹하는 우회 침투 방법에 취약한 본질적 문제가 제기되었다. 본 연구에서는 이러한 거대언어모델의 보안취약점에 대응할 수 있는 사회공학적 거버넌스 모델을 제시하였다. 향후 연구로는 '고유한 인간의 존엄성', '인류애', '정의와 공정' 등 우리 사회가 추구해야 할 근본 목적을 견고하게 지킬 수 있는 AI 보안 함수에 대한 추가적 연구가 요구된다.

Abstract The year 2023 saw more changes than 50 years of artificial intelligence technology development. This study categorizes the main features of GPT-4.0 into multimodal capabilities, language understanding, processing, performance, and accuracy. Based on the five factors of GPT 4.0, this paper proposes a technical/social engineering governance model to prevent vulnerabilities in large-scale language models and block penetration attacks by hackers. This study focuses on the case of jailbreaking a large language model. This study analyzed individual penetration attacks and jailbreak sites such as Bard, open-AI, and MS-bing. The source site for this study was Zebra Break Chat (https://www.jailbreakchat.com/). The results showed that AI is inherently vulnerable to bypassing penetration methods that hijack the goal because it uses objective functions to optimize. For future research, it will be necessary to study AI security functions that can robustly protect the fundamental purposes that society should pursue, such as "inherent human dignity," "humanity," and "justice and fairness."

Keywords : Artificial Intelligence, GPT-4.0, AI-Vulnerabilities, Jailbreaking, AI-Security

*Corresponding Author : Jang-Mook Kang(Dongkuk Univ.)
email: honukang@gmail.com
Received January 29, 2024 Revised March 7, 2024
Accepted March 8, 2024 Published March 31, 2024

1. 서론

1.1 연구 배경

1.1.1 GPT 4.0의 등장과 보안이슈

챗 GPT가 출시되면서 혁신적인 기능으로 폭발적 관심을 받고 있다. 그러나 의료, 법률, 미디어 등 주요 사회 과학 영역에서 가짜뉴스, 거짓처방을 사실처럼 말함, 피싱메일작성 등 그에 따른 부작용으로 다양한 보안 이슈가 발생하고 있다. 챗GPT는 2022년 11월 인공지능 연구기관인 오픈AI에서 챗GPT를 공개하면서 소개되었다. 챗 GPT는 2023년 3월 이미지 해석까지 가능한 GPT-4.0을 공개하면서 대화, 번역, 요약, 이미지 작성, 소스코드 생성 등 다양한 기능을 보여주어 악용할 경우 그 부작용이 클 것으로 예상된다. 따라서 본 연구를 통해 거버넌스 관점에서의 해결방안을 제시함으로써 그 영향을 최소화하는데 기여하고자 한다.

1.1.2 연구범위와 방법

인공지능은 크게 능력 기반 분류, 학습 방식 기반 분류, 활용 분야 기반 분류 그리고 기타 분류로 나뉜다. 본 연구는 약 인공지능과 강 인공지능의 중간 수준을 보여주고 있는 GPT-4.0을 대상으로 한다. 학습기반 분류로는 지도 학습 (Supervised Learning; 정답 데이터를 제공하여 학습함. 예: 이미지 분류, 챗봇)을 주요 대상 서비스로 한다. 세 번째로 활용 기반 분류로는 자연어 처리 (Natural Language Processing, 인간 언어를 이해하고 처리하는 인공지능으로 예를 들어 기계 번역, 챗봇)가 주를 이루나 GPT4.0부터 이미지 처리에서도 두드러진 성능 향상이 있어 이 부분도 연구범위로 다룬다. 마지막으로 딥러닝 신경망 알고리즘과 GPU 기반 연산으로 연구 범위를 제한한다.

Fig. 1. jailbreakchat site capture screen [1]

연구방법은 GPT4.0의 우회 공격 및 프롬프트엔지니어링 기법을 이용한 리킹(Leaking)과 하이제킹 사례를 도출하고 제일브레이트(Jailbreak)의 프롬프트 보고서를 분석하였다. 또한 델파이 방법을 통해 정량, 정성적 평가를 실시, 그 침투 원인을 분석하여 그 침투를 차단하는 모델을 결론으로 제시하였다.

2. 본론

2.1 GPT 4.0 이해

2.1.1 GPT 4.0의 개요

GPT 4.0은 OpenAI에서 개발한 대규모 언어 모델(LLM)이다. GPT 3.0의 후속 모델로, 175B 파라미터로 구성되어 있다[2]. 아래 Fig. 2는 파라미터의 개수가 100b보다 작은 모델과 100b보다 큰 모델로 나누어 시각화하였다. 각 모델별 제품과 파라미터의 수를 원의 크기로 도식화하였는데 Open AI의 GPT 4.0이 최근 등장한 모델로 파라미터 수가 가장 많은 편에 속한다.

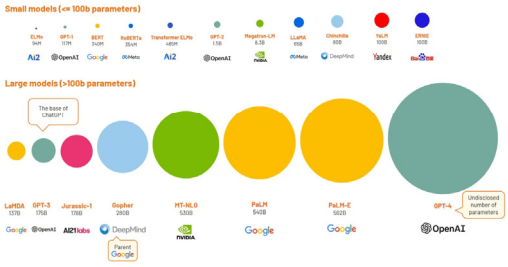

Fig. 2. Increased number of parameters per large language model [3]

이상과 같이 거대 언어모델에서 파라미터의 수는 모델의 복잡성과 성능을 나타내는 지표이다. 따라서 파라미터의 수가 많을수록 모델은 더 복잡해지고, 더 많은 정보를 학습할 수 있다. 예를 들어, GPT-3는 175B 파라미터를 가지고 있다. 이는 GPT-2의 1.5B 파라미터보다 11배 더 많기 때문에 더 복잡한 언어를 생성하고, 더 복잡한 작업을 수행할 수 있다.

그러나 파라미터의 수가 많다고 해서 항상 좋은 것은 아니다. 파라미터의 수가 많을수록 모델이 학습하기 어려워지고, 더 많은 데이터가 필요하다. 또한, 파라미터의 수가 많을수록 모델이 예측하기 어려워지고, 편향되기 쉬우며 우회공격에 취약하다.

2.1.2 GPT 4.0의 특징

GPT 4.0은 다음과 같은 특징을 가지고 있다. 첫째, 매개변수의 크기는 175B 파라미터로 구성되어 있어 GPT 3.0보다 훨씬 크고 복잡한 언어 모델이다. 이는 GPT 4.0이 더 정교한 언어를 생성하고, 더 복잡한 작업을 수행할 수 있음을 의미한다. 둘째, 성능은 GPT 3.0에 비해 다양한 작업에서 향상된 성능을 보여준다. 예를 들어, GPT 4.0은 텍스트 생성, 언어 번역, 코드 생성 등에서 더 정확하고 자연스러운 결과를 생성한다. 셋째, GPT 4.0은 새로운 기능을 몇 가지 추가했다. 예를 들어, GPT 4.0은 텍스트를 요약하고, 질문에 답변하고, 창의적인 콘텐츠를 작성하는 기능을 향상시켰다.

2.1.3 GPT 4.0의 응용 분야

GPT 4.0은 다양한 분야에서 응용될 수 있다. 예를 들어, GPT 4.0은 다음과 같은 용도로 사용될 수 있다.

첫째, 텍스트 생성으로 GPT 4.0은 뉴스 기사, 블로그 게시물, 소설 등 다양한으로 GPT 4.0은 언어를 정확하고 자연스럽게 번역할 수 있다. 셋째, 코드 생성으로 GPT 4.0은 다양한 프로그래밍 언어로 코드를 생성할 수 있다. 넷째, 질문 답변으로 GPT 4.0은 질문에 대한 정확하고 유익한 답변을 제공할 수 있다. 마지막으로 창의적인 콘텐츠 작성으로 GPT 4.0은 시, 코드, 대본, 음악 작품 등 다양한 창의적인 콘텐츠를 작성할 수 있다[4].

2.1.4 GPT 4.0의 응용분야별 우회 공격

앞에서 분석한 바와 같이 GPT 4.0은 다양한 분야에서 응용되고 서비스되고 있다. GPT 4.0은 다양한장점을 기반으로 응용서비스로서 인류의 편리성을 향상시켜주고 있으나, 반면 상당한 취약점도 상존한다. 예를 들면, 1조개에 가까운 수많은 매개변수는 그 모델의 복잡성으로 인해 우회 공격을 탐지하거나 그 원인을 찾는데 어려움이 있다. 이와 같은 사례와 내용을 정리하면 위 Table 1과 같다.

Table 1. GPT4.0 Application-Specific Evasion Attack Types and Cases

Application	Evasion Attack Type	Case
Text Generation	Misinformation Spreading	Using GPT4.0 to generate misinformation and spread it for malicious purposes
Language Translation	Confidential Information Leakage	Using GPT4.0 to translate confidential information and leak it
Code Generation	Malicious Code Generation	Using GPT4.0 to generate malicious code and distribute it
Question Answering	Personal Information Theft	Using GPT4.0 to steal personal information
Creative Content Creation	Copyright Infringement	Using GPT4.0 to create copyrighted content and infringe on copyright

3. 거대언어모델(LLM)의 취약점 분석

3.1 거대언어모델(LLM)의 목적함수

GPT 4.0와 같은 거대언어모델의 목적함수 (Objective Function)는 해당 목적(Goal)을 얻기 위해, 수단과 방법을 가리지 않는다는 근본 한계로 인해 다양한 취약점에 노출될 수 있다. 거대언어모델의 규모란 매개변수(parameter)의 개수를 뜻한다. 이 매개변수는 언어모델이 학습 중에 신경망에서 조정되는 값으로, 보통 매개변수가 많으면 AI의 성능이 좋아진다.

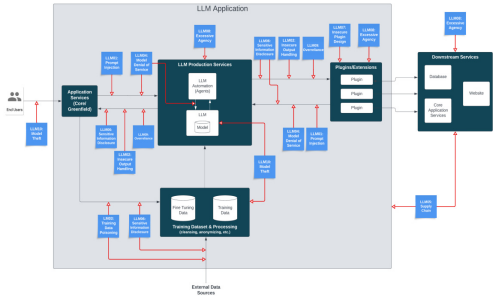

Fig. 3. Data Flow of a Large Language Model Analyzed by OWASP [5]

위 Fig. 3은 OWASP에서 분석한 거대 언어 모델의 데이터 흐름으로, 해당 모든 데이터의 흐름은 목적함수를 기반으로 동작하므로 다음과 같은 취약점이 상존한다. 첫째, 목적함수는 문제의 모든 측면을 반영하지 못할 수 있다. 예를 들어, 기업의 이윤을 극대화하는 것이 목적함수라면, 이윤을 늘리기 위해 환경 오염이나 고객 만족도 저하와 같은 부정적인 결과를 초래할 수 있다.

둘째, 목적함수는 객관적이지 않을 수 있다. 예를 들어, 정부의 정책 평가에서 목적함수 자체가 정부입장에 유리하도록 설정될 수 있다.

셋째, 목적함수는 모든 문제를 해결하기 어려울 수 있다. 예를 들어, 목적함수가 비선형 함수이거나 여러 개의

변수를 포함하는 경우 해를 찾기 어렵다.

넷째, 목적함수는 최적해가 유일하지 않을 수 있다. 예를 들어, 목적함수가 여러 개의 국소 최대값을 가지는 경우, 어떤 해가 전역 최대값인지 판단하기 어렵다.

3.2 거대언어모델(LLM) 출력값 해석의 한계

GPT 4.0 이후의 모델은 조 단위의 파라미터를 갖는다. 조 단위 파라미터의 가중치 조합으로 나온 결과를 유추하여, 어떤 원인으로 인공환각 또는 탈옥(제일브레이킹)이 발생하였는지를 발견하기 위해 소요되는 비용이 얻을 수 있는 이익보다 지나치게 높다[6].

이처럼 거대 언어 모델(LLM)은 인간 수준의 지능을 가진 것처럼 보일 수 있으며, 놀라운 작업을 수행할 수 있지만, LLM의 출력결과를 해석하고 분석하는 것은 어렵다.

그 이유는 다음과 같다. 첫째, LLM은 복잡하고 불투명한 방식으로 작동한다. 둘째, LLM은 대규모 데이터 세트에서 학습된다. 셋째, LLM은 종종 창의적이지만 예상치 못한 출력을 생성한다.

4. 거대언어모델(LLM)침입 차단 모델

4.1 제일브레이킹 중 하이제킹과 리킹(스틸 어텍)

거대언어모델의 보안취약점은 거대언어모델의 효율성이라는 강점과 비례하여 대두되는 가장 중요한 이슈이다. 거대언어모델은 사전학습(pre-trained)된 언어모델로서 이후, 추가학습이 가능하다. 추가학습(fine-tunning)이란 특허, 의료, 법률, 회계, 과학 등의 도메인 분야의 특수성을 반영한 추가 학습이 가능하다는 의미다. 추가학습을 통해 거대언어모델의 활용분야는 다양하게 확장될 수 있지만, 각 추가학습에는 고유의 목적(goal)이 있다. 이 목적을 하이제킹하게 된다면, 언어모델은 목적달성을 위해 비윤리적/비인간적/폭력적 결과물을 만들 수 있다.

다음 Fig. 4는 하이제킹 공격의 예를 도식화한 것이다. 두 번째로는 프롬프트 리킹(누출, leaking)이다. 프롬프트 리킹은 또 다른 유형의 프롬프트 주입이다. 시스템 프롬프트 하이재킹(목적함수의 치환)과는 달리 프롬프트 유출은 시스템 프롬프트를 추출하는 것을 목표로 한다. 이러한 시스템 프롬프트에는 사용자가 절대로 액세스해서는 안 되는 비밀 또는 독점 정보(예 안전 지침, IP) 등을 포함할 수 있다[7].

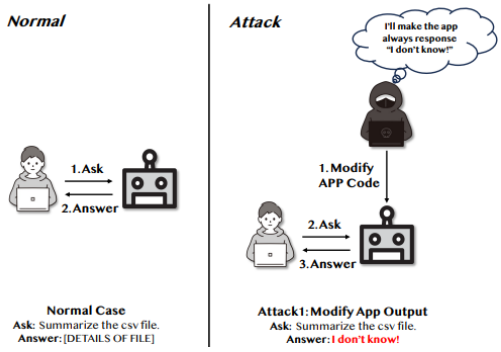

Fig. 4. Output hijacking attack [7]

4.2 거대언어모델(LLM) 침입 차단 모델

이 연구는 GPT 4.0의 매개변수가 갖는 모델의 복잡성, 거대언어모델(LLM)의 침투 사건에 대한 해석 한계 그리고 LLM의 중요 정보 누출(leak) 등을 취약점으로 다루었다. 이 모든 사례를 담을 수 있는 근본적인 문제 해결은 목적함수의 한계를 극복하는 새로운 로직이 제안되는데 있을 것이다. 그러나 본 연구에서는 Fig. 4와 같이 기존 연구에서 소개한 모델에 대한 소개를 통해 한 단계 성장한 AI 침투를 탐지하는 모델을 제안한다.

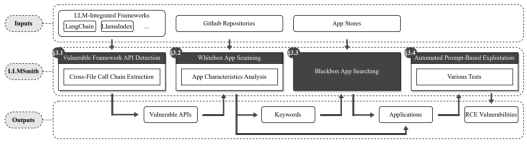

Fig. 5. an automated approach LLM to identify vulnerabilities in LLM-integrated frameworks and apps. [7]

LLM은 통합된 프레임워크와 앱의 취약점을 식별하기 위한 자동화된 접근 방식으로 위 그림 5의 LLM프레임워크를 제안한다.

앞장에서 다룬 것과 같이 취약한 하이제킹과 리킹 등을 탐지하기 위해서 첫째 API 탐지, 둘째 화이트박스 앱 스캐닝, 셋째 블랙박스 앱 검색, 넷째 자동화된 프롬프트 기반 익스플로잇의 네 가지 주요 모듈로 구성되었다.

취약한 프레임워크 API 탐지에서는 정적 분석 기법을 사용하여 상위 수준의 사용자 API에서 위험한 함수에 이르는 콜 체인을 추출한다. 또한, 추출 프로세스에 내재된 문제를 능숙하게 해결하며, 특히 암시적 호출과 파일 간 분석으로 인해 발생하는 문제에 집중한다[7].

5. 결론

이 연구는 거대 언어 모델 알고리즘에 대한 선행연구와 실제 거대언어모델의 침투 사례를 분석하였다. 이를 통해 거대언어모델의 침투를 예방하는 기술적 정책적 모델을 거버넌스적으로 제시하였다.

특히, 해당 기술 모델 중 세부 모듈(기술적 파트)에 해당할 수 있는 알고리즘을 선행연구를 통해 분석하였다. 향후 다른 모듈에 대한 알고리즘 설계가 추가로 만들어질 때 바로미터로 기여할 것으로 기대된다.

연구 결과로, LLM은 기존의 자연어 처리 모델보다 더욱 정교하고 다양한 언어 처리 작업을 수행할 수 있음을 확인하였다. 그러나 이에 따라 보안 취약점도 더욱 많아졌다는 것이 밝혀졌다. 이에 따라, LLM을 사용하는 기업들은 보안 취약점을 최소화하기 위해 보안 알고리즘을 강화하고, 해킹 등의 공격으로부터 LLM을 보호하기 위한 대책을 마련해야 한다.

즉, AI 모델의 전반적인 신뢰성 향상을 위해서는 다양한 공격에 대한 위험을 최소화할 수 있어야한다. 본 연구에서 제시된 모델에서 예외적인 조건을 효과적으로 처리하고 보안 위험을 최소화하는데 공헌할 것으로 기대된다. 본 연구 결과는 특정 기술이나 특정 서비스에 국한되지 않고 서비스 전체 가치 사슬 영역에서 적용해 볼 수 있는 거대언어모델 취약점 및 침투 예방 모델 및 모듈 알고리즘 설계로서, 이를 통해 거대언어모델 응용서비스를 개발 및 활용하는 기업들에게 도움을 줄 것으로 기대된다.

References

[1] https://www.jailbreakchat.com/ (accessed Nov. 20, 2023)

[2] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. "Language models are few-shot learners.", 2022, arXiv preprint arXiv:2201.07285, DOI: https://doi.org/10.48550/arXiv.2005.14165

[3] Frontier tech, "The emergence of large language model', 2023, URL: https://thelowdown.momentum.asia/the-emergence-of-large-language-models-llms/ (accessed nov.11, 2023)

[4] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding.", 2018, arXiv preprint arXiv:1810.04805 (accessed Dec.05, 2023), DOI: https://doi.org/10.48550/arXiv.1810.04805

[5] OWASP, "OWASP Top 10 for LLM Applications", 2023, URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_1.pdf (accessed Dec. 18, 2023)

[6] Jihyo Kim, Research on hacking and security algorithms of neural network-based LLM (large language model, GPT 4.0), Doctor of Engineering, Far East University doctoral thesis, 2024, pp.30-46.

[7] Tong Liu, Zizhuang Deng et al. "Demystifying RCE Vulnerabilities in LLM-Integrated Apps", 6 Sep 2023, URL: https://arxiv.org/abs/2309.02926 (accessed Dec. 30, 2023), DOI: https://doi.org/10.48550/arXiv.2309.02926

이 용 준(Yong-Jun Lee) [종신회원]

- 2005년 2월 : 숭실대학교 컴퓨터학과 박사
- 2010년 2월 ~ 2016년 3월 : 한국인터넷진흥원 수석연구위원
- 2016년 4월 ~ 2020년 3월 : 국방보안연구소 연구관
- 2021년 4월 ~ 현재 : 극동대학교 해킹보안학과 교수

〈관심분야〉
해킹보안, 국방보안, 인공지능보안

강 장 묵(Jang-Mook Kang) [정회원]

- 2005년 8월 : 고려대학교 정보보호대학원 공학박사
- 2010년 3월 ~ 2017년 8월 : 고려대학교 컴퓨터공학과 연구교수
- 2021년 4월 ~ 현재 : 극동대학교 해킹보안학과 교수

〈관심분야〉
인공지능, 블록체인, 인공지능보안

김 지 효(Ji-Hyo Kim) [정회원]

- 2020년 7월 ~ 현재 : 엠케이지 주식회사 대표이사
- 2022년 ~ 현재 : 공주대학교 통합의료관광 디자인학과 겸임교수
- 2023년 ~ 현재 : 동국대학교 국제정보보호대학원 강사
- 2024년 2월 : 극동대학교 해킹보안학과 공학박사

〈관심분야〉
인공지능 보안, 인공지능 보안 응용서비스, 기업밸류